BMC Genomics

BioMed Central

Software

Open Access

ProSeeK: A web server for MLPA probe design

Lorena Pantano, Lluís Armengol, Sergi Villatoro and Xavier Estivill*

Address: Genes and Disease Program, Center for Genomic Regulation (CRG), Doctor Aiguader, 88, 08003 Barcelona, Catalonia, Spain

Email: Lorena Pantano - lorena.pantano@crg.es; Lluís Armengol - lluis.armengol@crg.es; Sergi Villatoro - sergi.villatoro@crg.es; Xavier Estivill* - xavier.estivill@crg.es

* Corresponding author

Published: 28 November 2008

BMC Genomics 2008, **9**:573 doi:10.1186/1471-2164-9-573

Received: 4 June 2008

Accepted: 28 November 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/573

Abstract

Background: The technological evolution of platforms for detecting genome-wide copy number imbalances has allowed the discovery of an unexpected amount of human sequence that is variable in copy number among individuals. This type of human variation can make an important contribution to human diversity and disease susceptibility. Multiplex Ligation-dependent Probe Amplification (MLPA) is a targeted method to assess copy number differences for up to 40 genomic loci in one single experiment. Although specific MLPA assays can be ordered from MRC-Holland (the proprietary company of the MLPA technology), custom designs are also developed in many laboratories worldwide. After our own experience, an important drawback of custom MLPA assays is the time spent during the design of the specific oligonucleotides that are used as probes. Due to the large number of probes included in a single assay, a number of restrictions need to be met in order to maximize specificity and to increase success likelihood.

Results: We have developed a web tool for facilitating and optimising custom probe design for MLPA experiments. The algorithm only requires the target sequence in FASTA format and a set of parameters, that are provided by the user according to each specific MLPA assay, to identify the best probes inside the given region.

Conclusion: To our knowledge, this is the first available tool for optimizing custom probe design of MLPA assays. The ease-of-use and speed of the algorithm dramatically reduces the turn around time of probe design. ProSeeK will become a useful tool for all laboratories that are currently using MLPA in their research projects for CNV studies.

Background

The technological evolution of platforms for assessing genome-wide copy number imbalances [1] has allowed the discovery of an unexpected amount of human sequence involved in duplications and deletions (termed copy number variants or CNVs). In terms of sequence coverage, this is the most important type of human variation identified so far and can make an important contribution to human diversity and disease susceptibility (see [2] for review). So far, derived from the study of several hundreds of individual genomes, ~19% of the euchromatic portion of the human genome has been reported as variable (mainly in copy number) [3]. Several studies have shown the relationship between CNVs and disease phenotypes [4,5].

MLPA [6], Multiplex Ligation-dependent Probe Amplification, is a targeted method to assess copy-number differ-

ences for up to 40 genomic regions in one single experiment. Each MLPA probe is composed of two oligonucleotides that are only ligated, and subsequently amplified, if specifically hybridized to the target locus. The left probe oligonucleotide (LPO) is made of a complementary sequence of an universal forward PCR primer at its 5' end, plus the specific hybridizing sequence (LHS) at its 3' end. The right oligonucleotide (RPO) has the specific hybridizing sequence (RHS) at its 5' end followed by the complementary sequence to the reverse universal PCR primer, at the 3' end. After ligation all probes are amplified, by means of a universal primer pair, in a multiplex PCR reaction. This PCR produces loci-specific amplicons due to a stuffer sequence located between the hybridizing and the universal sequences. They are then resolved by capillary electrophoresis and copy number of each region is measured as a function of peak intensities of the MLPA amplification products (Figure 1).

Although specific MLPA assays can be ordered from MRC-Holland (the proprietary company of the MLPA technology), custom designs are also developed in many laboratories worldwide. After our own experience, an important drawback of custom MLPA assays is the time spent during the design of the specific oligonucleotides. Due to the large number of probes included in a single assay, a number of restrictions need to be met in order to maximize specificity and to increase success likelihood. Given the tedious stepwise procedure that is followed, the goal of ProSeeK is to automate the process of probe design and to obtain the best candidate probes for a given region.

Implementation

ProSeeK is presented as an easy-to-use and point-and-click web interface. Is implemented in CGI (Common Gateway Interface) Perl scripts and made accessible to the user using PHP on top of an APACHE server with MYSQL database support. It is accessible through the Internet (at http://davinci.crg.es/estivill_lab/mlpa) with IE5.0 and Netscape 7 or higher, from any platform. By making use of universally available web GUIs, the system solves the problem of portability of this software. No client-side software installation is required.

The algorithm for probe design consists of several modules (Figure 2) that are run iteratively. (1) Sequence Checker, ensures that a valid sequence format is entered by the user. (2) Hybridization Finder, that identifies a set of hybridizing sequences (HSs), with the correct size, that are required to start and end with either a C or a G (according to the MRC-Holland protocol advises). Candidate HSs are filtered based on melting temperature and GC content, according to the set of thresholds provided by the user, and subsequently added primers and stuffers as needed. (3) Sequence Aligner, that performs a genome alignment using BLAT [7] to identify the optimal HS. Candidate HSs, regardless of having single or multiple matches to the genome, are filtered by the e-value of the alignment before passing to the next module. In the case that the HSs map onto a copy number variable (CNV) region or segmental duplication (SD) in the reference genome, the HSs are only recognized as optimal if the multiple matches are perfect and the other possible matches are below the e-value threshold. In the case that the probes designed are located in CNV or in SDs regions, this information is shown to the user in the output flagging them as 'CNR' and 'SD' respectively. (4) HS Trimmer, conveniently splits the HSs to fulfill user-entered criteria in terms of length, melting temperature and global sequence composition. (5) Results Generator, presents results to the user. ProSeeK can be asked to retrieve different results: the "partial" will only produce the optimal design for the left and right hybridizing sequences, while the "complete" will produce the whole oligonucleotide sequences corresponding to the LPO and RPO (see above). (6) Data Keeper, takes care of storing the results in a personal space of the database for future retrieval of the designs.

Input to server

ProSeek requires the DNA sequence of the target region in which the MLPA probes will be designed. Several parameters can be used to restrict the probe design: (1) maximum GC content, (2) maximum melting temperature (Tm) of the hybridizing sequence, (3) Blat e-value (minimal length that the Blat will detect as a match), (4) hybridizing sequences (HS) length, (5) stuffer sequence, (6) sequence of the universal primers to flank the HS, and (7) desired probe length. (Additional file 1).

Output from server

After computing all available possibilities, ProSeek produces a table in HTML format containing optimal probes which are presented to the user, together with their characteristics, which include position within the user-entered sequence, genome mapping, GC content, melting temperature, probe sequence, nucleotide length, self-folding capacity (i.e. DNA secondary structure prediction using DINAMelt Server [8]), and links to the UCSC Genome Browser [9] and to the Database of Genomic Variants [10] The projects are kept on the server for one month, so the users can retrieve their results at any time by returning to the website and identifying himself on the initial web page. (Additional file 1).

Conclusion

A number of high-throughput technologies have become available to address the genome-wide detection of structural variations in humans. An important drawback of these new methods is that a huge amount of false positive

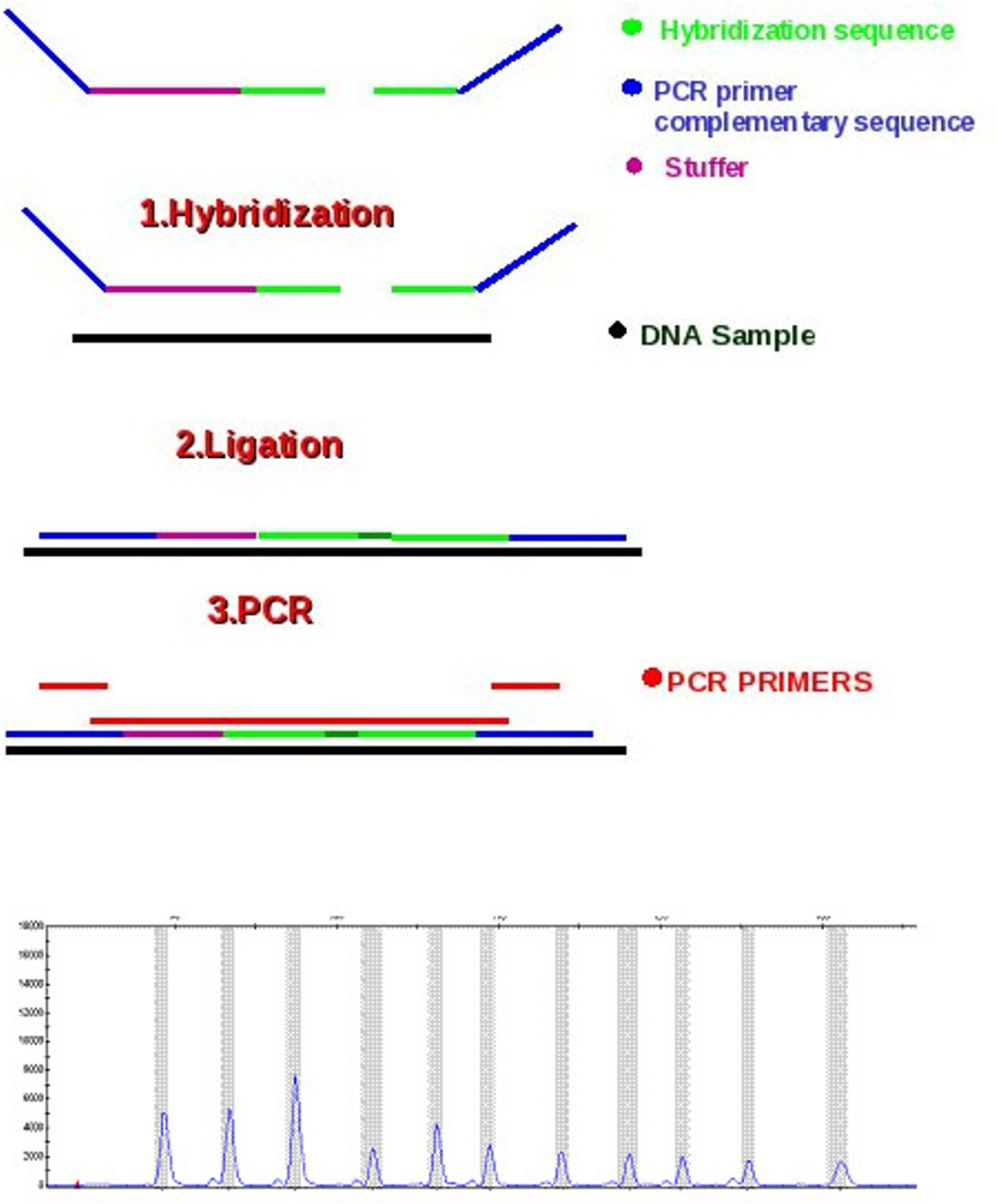

Figure 1
MLPA assay. Typical steps in a MLPA assay and the final output where each peak represents a probe in the experiment.

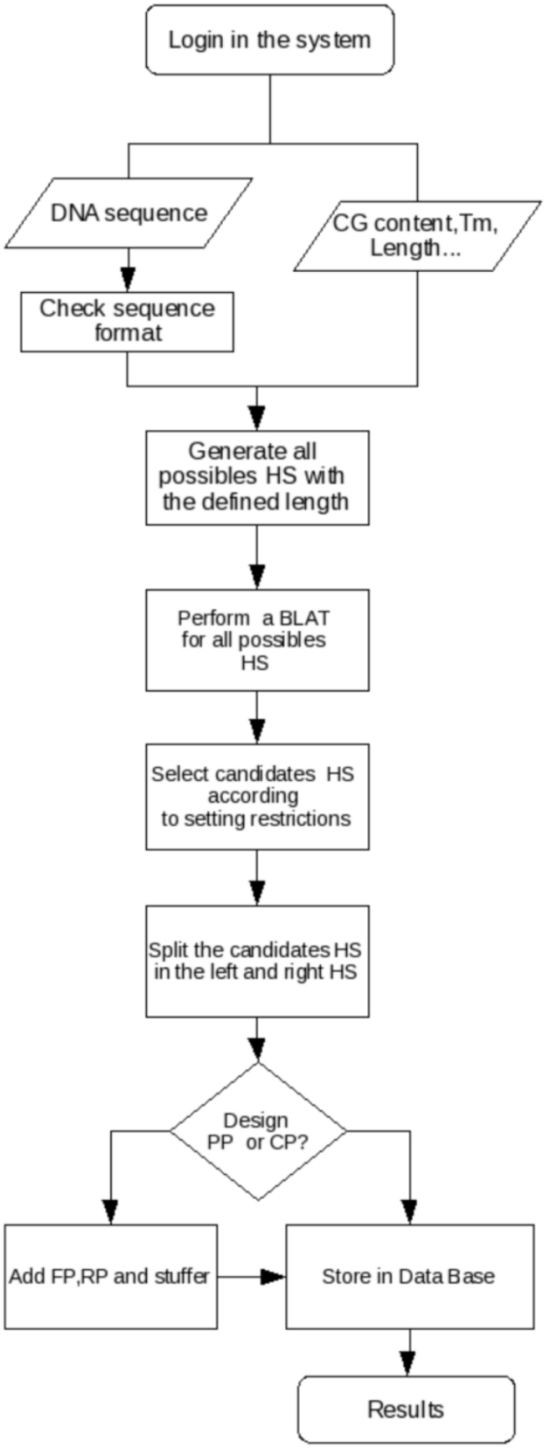

Figure 2
Protocol. Flowchart describing the implementation of ProSeek algorithm. HS (Hybridizing sequence), PP (Partial project), CP (complete project), FR (Forward primers), RP (Reverse primers).

results typically arise after analysis, thus it is mandatory to validate observations made with these technologies using alternative and more reliable approaches. Among others, due to its simplicity, robustness and relative low price, the MLPA is often used as a targeted method to assess copy-number differences. One important inconvenience is the required time for designing the probe-mixes to target the desired regions, since a lot of restriction should be fulfilled to get a sensitive, specific and reproducible experiment. To overcome this aspect, we developed ProSeek that produces the optimal probes for the regions of interest. ProSeeK is, to our knowledge, the first algorithm for the design of MLPA probes, that allows saving time and improving accuracy of MLPA assays.

Availability and requirements

- Project name: ProSeeK
- Project home page: http://davinci.crg.es/estivill_lab/mlpa
- Programming language: Perl
- License: GNU General Public License

Authors' contributions

LP initiated this web server project, wrote the original source code, constructed the web interface, implemented it on the server and wrote the manuscript. LA conceived the server and participated in manuscript writing. XE revised and helped to write the manuscript. SV helped to design the web interface particularly from the viewpoint of an experimental research field. All authors contributed to the final manuscript and agreed the final version.

Additional material

Additional file 1

Tutorial. *A complete tutorial explaining step by step the ProSeeK procedure.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-573-S1.pdf]

Acknowledgements

This work has been supported by Spanish Ministry of Science and Innovation under NOVADIS project (SAF2008-00357), and by the European Commission under Aneuploidy (037627) and ENGAGE (201413) projects.

References

1. Carter NP: **Methods and strategies for analyzing copy number variation using DNA microarrays.** *Nat Genet* 2007, **39:**S16-S21.
2. Feuk L, Carson A, Scherer S: **Structural variation in the human genome.** *Nature Reviews* 2006, **7:**85-97.

3. Scherer W, Charles L, Ewan B, Altshuler D, Eichler E, Carte N, Hurles M, Feuk L: **Challenges and standards in integrating surveys of structural variation.** *Nat Genet* 2007, **39(7 Suppl):**S7-S15.
4. Feuk L, Marshall C, Wintle R, Scherer S: **Structural variants: changing the landscape of chromosomes and design of diseases studies.** *Hum Mol Genet* 2006, **15:**R57-R66.
5. McCarroll S, Altshuler D: **Copy-number variation and association studies of human disease.** *Nat Genet* 2007, **39:**S37-S42.
6. Schouten J, McElgunn C, Waaijer R, Zwijnenburg D, Diepvens F, Pals G: **Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification.** *Nucleic Acids Res* 2002, **30:**e57.
7. Kent W: **BLAT – The BLAST-Like Alignment Tool.** *Genome Res* 2002, **12(4):**656-664.
8. Markham NR, Zuker M: **DINAMelt web server for nucleic acid melting prediction.** *Nucleic Acids Res* 2005, **33:**W577-W581.
9. Kent W, Sugnet C, Furey T, Roskin K, Pringle T, Zahler A, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
10. Iafrate A, Feuk L, Rivera M, Listewnik M, Donahoe P, Qi Y, Scherer S, Lee C: **Detection of large-scale variation in the human genome.** *Nat Genet* 2004, **36:**949-51.